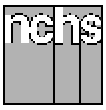
Sample Design, Weighting, and Variance Estimation for the new NHIS

J. Neil Russell, Ph.D.



CDC
CENTERS FOR DISEASE CONTROL AND PREVENTION

Topics covered

- **Sample design (1995 - 2004)**
- **Weighting**
- **Variance estimation**

NHIS - Sample Design

Nationally representative sample of households

- **Civilian non-institutionalized population**
- **Probability sample**
- **Multistage, stratified, cluster design**

NHIS-Target Population

Excludes persons in:

- **The military**
- **Long-term care facilities**
- **Prison**

Eligible for interview:

- **Group homes**
- **Apartment buildings**
- **Campuses**

NHIS-Sample Features

Probability Sample - Each household has a known probability of selection

NHIS-Sample Features

Multistage

- **Sample selection done in stages**
- **Each stage uses a different frame**

Stratified

- **Sampling units sorted by characteristics of interest**

NHIS-Sample Features

Cluster

- **Select groups of households near each other**
- **More efficient; less costly**

Multistage Stratified Design (1)

Divide US into 1,995 Primary Sampling Units (PSUs)

- **PSUs stratified by state**
- **PSUs stratified within State by MSA/non-MSA**

Multistage Stratified Design (2)

1st stage – select 358 PSUs

- Largest 95 PSUs always selected (Self-Representing)
- Using PPS, 263 NSR PSUs selected (Nonself-Representing)
- Usually two PSUs per state

Multistage Stratified Design (3)

2nd stage –

- Within each PSU, stratify all blocks by race/ethnicity density
- Then, systematically select clusters of blocks (Secondary Sampling Units - SSU)

Multistage Stratified Design (4)

3rd stage –

- **Within each block group (SSU), select clusters of 4-12 households**
- **Interviewers list all addresses**

NHIS - Oversample

- **Sample SSUs in areas with larger minority populations at higher rate**
- **Retain all households with Black or Hispanic member**
- **Retain only a portion of other households**

Sample Design Summary

- Representative of US and 4 regions
- Household is the sampling unit
- Complex design
- Over-sample of Black and Hispanic households

Analytic issues: Intro

- Weights are needed
  - What are “weights”?
  - Why should they be used?
- Statistical procedures must take complex sample design into account
  - Why special procedures for variance estimation?

NHIS Weights

- **“Weights” exist on data files as a variable**
- **A weight inflates each observation**
- **Weights are unequal**
- **Weights are adjusted**
- **Sum of weights = national total**

Weights: product of four components

- **Probability of selection**
- **Non-response adjustment**
- **First-stage ratio adjustment**
- **Second-stage ratio adjustment (post-stratification)**

Probability of selection

- **Units multiplied by inverse of selection probabilities, based on:**
  - **PSU**
  - **Segment (SSU)**
  - **Household**

Non-response adjustment

- **Weight inflated to account for non-interview units**
- **Assumes responding households represent non-responding households**

1st stage ratio adjustment

- Weight adjusted for non-self representing (NSR) PSUs based on:
  - MSA / non-MSA
  - Race / ethnicity
    - Hispanic
    - Non-Hispanic Black
    - Non-Hispanic Other
  - Region (NE, MW, S, W)

2nd stage ratio adjustment

- Weights adjusted to Census population control totals based on:
  - Age
  - Sex
  - Race / ethnicity
    - Hispanic
    - Non-Hispanic Black
    - Non-Hispanic Other

Why use weights?

- To produce national estimates
- To produce unbiased estimates

Weights, con't

- Failure to use weights –
  - Totals affected
  - Means and proportions distorted
  - Certain estimators biased

1997: Percent Race/ethnicity

Race/ ethnicity	Unweighted	Weighted
Hispanic	21.2	11.2
NH White	60.8	72.2
NH Black	14.2	12.2
NH Other	3.9	4.3

Source: 1997 NHIS

Which weight to use?

- Person level files
- Appropriate weight: WTFA
- Can be used on almost all variables

Variance estimation

- **What is variance and standard error?**
- **How does sample design affect variance?**
- **Why use special procedures to calculate variance?**

Variance

- **What is variance?**

An "average" distance of all data points distributed around the mean. Or, a measure of the spread of the data.

$$\text{var.} = \frac{\sum (X - \bar{X})^2}{n - 1}$$

Variance

- **What is standard error?**

  Another measure of the spread of the data. It is also a "proxy" for sampling error.

$$\text{stan. err.} = \sqrt{\text{var.}}$$

Variance

- **Why does variance matter?**

$$\textbf{t - statistic} = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\text{var}_1 + \text{var}_2}}$$

Variance

Or, why does standard error matter?

Confidence interval (C.I.)

C.I. = $\bar{X} \pm 1.96$ (stan. err.)

T-test: M vs. F on Health Status

Stat.	SRS Unweighted		SRS Weighted		Complex Weighted	
	Male	Female	Male	Female	Male	Female
Mean	2.023	2.133	1.992	2.091	1.992	2.091
S.E. Mean	.00462	.00457	.00008	.00009	.0068	.0066
T-stat	-16.93		-775.96		-16.52	
d.f.	102,937		265,000,000		339	

Source: 1997 NHIS

NHIS Sample design

- Probability (known selection probability)
- Stratified (State and MSA / non-MSA)
- Multistage (PSU, SSU, blocks, housing unit)
- Cluster (households within same block)

Sample design and variance

- Impact of complex survey design (compared to a Simple Random Sample - SRS)
  - Stratification - decreases variance
  - Multiple stages - increases variance
  - Clustering - increases variance

Sample design and variance

- **Sample design affects variance computation**
- **A complex survey will produce larger variances than a SRS**
- **Need a different method of variance calculation if sample is other than SRS**

Review: variance

- **NHIS has a complex survey design**
- **Must use appropriate statistical procedures to account for complex sample design**
- **Variance estimation would be biased is sample design is ignored**

Review: variance

- Most statistical software assume SRS
- Need computer software that can calculate variance under the assumptions of a complex survey design

Variance estimation, con't

- Computer software for variance estimation of complex surveys
  - Recommend SUDAAN
  - Other possibilities
    - EPI Info
    - STATA
    - SPSS add-on module
    - WestVar
    - SAS ver. 8.0